View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by PUblication MAnagement

The Lexical Grid: Lexical Resources in Language Infrastructures

Monica Monachini Claudia Soria
Nicoletta Calzolari

March 2008

Istituto di Linguistica Computazionale del CNR (ILC-CNR)
Area della Ricerca del CNR
Via Giuseppe Moruzzi N° 1
56124 Pisa
ITALY

Summary

1 Today's situation of Language Resources and Language Technologies

Language Resources are recognized as a central and strategic component for the development of any Human Language Technology system and application product. They play a critical role as horizontal technology and have been recognized in many occasions as a priority also by national and supra-national funding agencies, i.e. the European Commission (EC) and NSF. In Europe, the EC played an essential role in funding a number of initiatives (such as EAGLES, ISLE, ELRA) to establish some sort of coordination of LR activities, and a number of large LR creation projects, both in the written and in the speech areas. LRs have acquired larger and larger importance and impact in the last two decades, when more and more activities, both at the European level and worldwide, have contributed to substantial advances in knowledge and ability of how to represent, create, acquire, access, exploit, harmonize, tune, maintain, distribute LRs. Over the past two decades, the HLT community has invested substantial effort in the creation of computational (written and spoken) lexicons and compendia of semantic information (e.g. Wordnets, FrameNets, ontologies) together with (written and spoken) language corpora annotated for all varieties of linguistic features, which comprise the central resource for current NLP research.

However, after a few years of strong involvement from the EC as well as national funding agencies, for a while, new initiatives have emerged here and there in a rather opportunistic way and didn't have a thoroughly well-articulated long-term vision. This has lead to the creation of disjointed language resources and tools, which are often not reusable and/or interoperable.

Recently, there have been important signs of attention to the LR area again. From the industry side, there is a clear growing interest in the use of LRs, in particular for multilingual applications. A sign of the wide resonance of LRs can certainly be seen in the success of the Language Resources and Evaluation Conference (LREC), in the establishment of the new international journal *Language Resources and Evaluation* (Ide and Calzolari 2005) – both initiatives of ELRA – , and in the attention paid by ISO to the standardization of LRs at large. From the EU side, both the Call for a Thematic Network for Language Resources (the new born FlaReNet project) and the Call for a Research Infrastructure of Language Resources and Language Technologies for the Humanities and Social Sciences (CLARIN project) show a renewed awareness of LR strategic relevance, as the necessary *infrastructure* to develop a coherent and robust LT *platform* for accessing digital content.

1.1 The need for a Language Infrastructure

Most of the existing language data resources and NLP tools/systems have been created independently, resulting in a situation where data format, annotation scheme, access method and other features are all idiosyncratic. The huge amount and diversity of language resources and tools, strongly demands for a forward-looking view in order to try to weave the various resources scattered over different sites into a single organism of language repositories and services. This can only be achieved through a coordinated, community-wide effort that will ensure the contribution of the main actors from the various areas. Recent developments of the Semantic Web, progresses of the associated methodologies and the availability of mature standards for content interoperability suggest that time and circumstances are ripe for defining a new paradigm for language resources and technologies and setting up the basis of an open and distributed language infrastructure. An infrastructure of this sort is also expected to facilitate further development of language data resources and NLP functionalities. Infrastructure building is a time-consuming activity and only robustness and persistency of the offered solutions will convince researchers and users.

There are a number of initiatives and projects, where such notions are playing a prominent role and efforts in these directions are being carried out.

2 Language Infrastructure: one of the ILC missions

The integration and exploitation of language resources and tools into an architecture where users can combine elements of static language resources and dynamic processing resources is an active research topic being pursued at CNR-ILC, both independently and in the framework of international projects.

2.1 Language Grid

Our institute is a partner of the Language Grid[1] initiative, led by NICT, which aims at providing an open and distributed infrastructure on which existing language resources, NLP tools/systems, newly created community-based resources can be efficiently combined and language services can be effectively composed, delivered, and utilized (Ishida, 2006). This framework should be seen as a General Language Infrastructure, called GLI, is meant to accommodate language resources and technologies world-wide and enhance international collaboration and multicultural cooperation. More precisely, a GLI is an open and web-based software platform to which resources can be easily plugged in, and on which tailored language services[2] can be efficiently composed, disseminated and consumed (Hayashi et al, 2008a).

The assumption of the Language Grid project is that the infrastructure should be based on shared ontologies which cover all possible elements of a GLI – services and resources – in order to provide appropriate processes involving service discovery, planning and invocation in the form of advanced and efficient workflows able to combine atomic services into composite ones on the basis of end-user requirements. Ontology offers the possibility to have meta-descriptions of elements and gives a solid foundation. Language Grid proposes a high-level configuration of these ontologies, which are integrated into comprehensive language service ontology which incorporates not only processors, but also data such as lexicons and corpora, linguistic objects such as linguistic expressions, meanings, meaning description, and linguistic annotation from many perspectives.

ILC contributes to Language Grid objectives towards the ontology-driven infrastructure, mainly concentrating on two research topics:

- Development of the comprehensive language services ontology
- Demonstration of composite language services, particularly lexical services

Fulfilling these objectives and tasks, the definition of modern services for lexical resources, first imply the identification of requisites an integrated lexical resource platform should comply with. These requisites are lucidly described in Calzolari 2008 where the vision of the "Lexical Web", which motivates and orientates our contribution to Language Grid, is presented.

2.2 CLARIN

Language Grid shares issues of interoperability and reusability of language data resources and tools/systems with another project, CLARIN (where our research group is involved), even though their primary objectives are totally different. This calls for an opportunity to work out a common strategy for these crucial issues.
CLARIN[3] is an ESFRI project for the development of a pan-European integrated and interoperable infrastructure of language resources and technologies. Similarly to Language Grid, it aims at addressing the current fragmentation by offering a stable, persistent, accessible and extendable infrastructure. Different are target users, since CLARIN points to scholars of all disciplines, in particular of the humanities and social sciences. A strong preparatory phase is expected to pave the way towards the necessary maturity of such an infrastructure which should enable the development of "e-Humanities". The infrastructure will offer persistent and secure services and provide easy access to LRs and LTs: the user will have access to repositories of data with standardised descriptions, processing tools ready to operate on standardised data, and guidance from distributed knowledge centres. All this will be available on the web using a service oriented architecture based on secure grid technologies. CLARIN will turn existing, fragmented LRs and LTs into accessible, stable services that any user can share, adapt and repurpose, building upon the rich history of European and national initiatives.

2.3 FLaReNet

Another initiative where the definition of the scientific, organizational and economic conditions and winning strategies for the development of a modern language infrastructure will be addresses is the recently approved

[1] http://langrid.nict.go.jp
[2] Here a *language service* simply means a web service whose functionalities are generally related to human language; it can range from simple dictionary access to more complicated linguistic analysis, as well as translation.
[3] http://www.clarin.eu

Thematic Network *FLaReNet*[4]. FLaReNet will act as a forum to facilitate interaction among LR stakeholders with the objectives of re-creating the international cooperation of the LR community. FLaReNet departs from the assumption that LRs present various dimensions and must be approached from many angles: technical, but also organizational, economic, legal, political, also addressing multicultural and multilingual aspects (essential when facing access and use of digital content in today's Europe). FLaReNet will bring together leading experts (academic and industrial) to ensure coherence of LR-related efforts in Europe. Among the various aspects to be tackled, FLaReNet will discuss "… *new strategies to address the current fragmentation by offering a stable, persistent, accessible and extendable infrastructure that will enable the integration of so far partial solutions into broader architectures*, … *anticipating the needs of new types of LRs*…" and will try "*to convert existing and experimental technologies related to LRs into useful economic and societal benefits*". It is of utmost importance that the Language Grid project will go step by step with FlaReNet by delivering the results of its collaborative research initiative.

This joint effort will contribute to identify which pillars and new building blocks do emerge today encompassing the realisation of a comprehensive notion of a *distributed infrastructure for LRs*, a Language Grid.

3 Towards an Ontology-driven Lexical Infrastructure: the *Lexical Grid*

Mixing considerations on what is needed for a Language Infrastructure, issues relevant to the establishing a modern lexical resource infrastructure, a *Lexical Grid*, – undoubtedly a key part of the broader General Language Infrastructure – are touched here.

CNR-ILC contributions to the realization of the Language Grid ontological vision of a language infrastructure, first of all, consists the definition of the conditions for an ontology of lexical resources, their communication as well as their interaction protocols. These, as already mentioned above, are inspired by Calzolari 2008 and totally embrace the vision of a "Lexical Web" presented there. Moreover, the dimensions which are relevant to the creation of such a modern architecture and the driving forces are individuated in Calzolari 2007, as follows:

- *Interoperability*, and even more *content interoperability*: language is the key mediator to access content, knowledge, ontologies;
- *Collaborative creation and management of LRs* (even on the model of wiki initiatives)
- *Sharing of LRs*, as a new dimension of the distribution notion;
- *Dynamic LRs*, able to enrich themselves.

CNR-ILC also contributed to the definition of lexical service ontology and to the demonstration of Language Grid (technical) soundness. This has been done through a variety of approaches and some research activities presented below.

Definition of a lexical service ontology

i) standardization activities; it goes without saying that, in order to realize its ontological vision, the Language Grid initiative is building on the work done within the ISO TC37/SC4 committee on Language Resources management
ii) ontology for lexicons and lexicon services;
iii) special-domain lexical data bases;
iv) "ontologisation" of lexico-semantic resources;

Lexical services on a GLI

v) *lexical services* for automatic up-loading of lexical databases;
vi) experimental procedures for mapping/unifying existing lexicons;
vii) architectures for managing/merging/integrating lexical resources
viii) UIMA framework for resource and tool sharing and interoperability

[4] *Fostering Language Resources Network*, a Thematic Network proposed in the context of the last eContent*plus* Call, will be coordinated by CNR-ILC.

3.1 Standardization: the Heart of a Language Infrastructure

To address this issue, standardization is inevitable: standardized APIs are necessary for NLP tools/systems; standardized data semantics as well as data format are required for language data resources.

Big steps forward have been made with respect to standardization, which is among the necessary pre-conditions to integration and exploitation of language resources and tools into a same architecture where they can be combined. Standards are the precondition for content interoperability and for providing tools able to process data with standardized descriptions and return standardized output. Expectations of the scientific and industrial community about standards are that, once made operational in an integrated resource platform, they will be beneficial to the definition of both standardized access functions and automated workflows. The challenge for them is to enable a modern service-oriented infrastructure with a set of stable language services.

Standardization for lexical resources had been studied and developed by a series of projects like GENELEX, EAGLES, MULTEXT, PAROLE, SIMPLE and ISLE (www.ilc.cnr.it/EAGLES96) (Calzolari et al, 2003). However, although the standards issued by these projects had been widely adopted by research institutions and academy, they also needed adoption within the industrial community to support advanced language technologies for content access and sharing. In order to reach wide industrial audience, production and ratification by an official International body seemed necessary. ISO devoted much attention to the development of standards for developing and managing Language Resources, with Committee e Working Groups especially dedicated to various aspects.

3.2 The ISO Lexical Markup Framework

The ISO TC37/SC4 WG4 dedicated to NLP lexicons is in charge of defining lexical standards. The result is the LMF (Lexical Markup Framework) standard (Francopoulo et al, 2006). The design of a standard for lexicons poses a great challenge. Many of them are complex and very different in nature from each other, because they contain different types of information. In order to avoid having a one block specification that is difficult to understand, LMF adopted a modular organization. As a consequence, LMF (http:// lirics.loria.fr/ documents.html) is made up of a core model, a sort of simple skeleton, and various semi-independent packages of notions, used for the various linguistic layers that make up a lexicon. They can be combined together as needed to meet different requirements and describe an LMF-conformant lexicon.

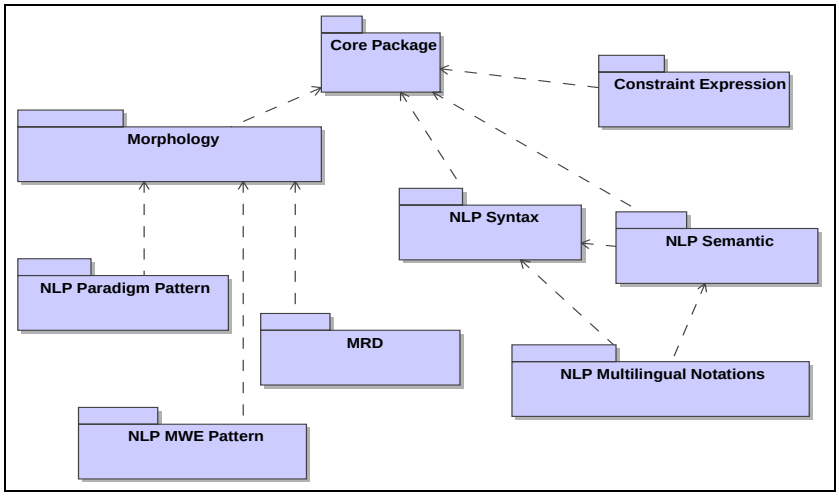
The reasons behind this choice are that the lexicon is not to be considered as an 'ivory tower'. A lexicon is a repository to be correctly integrated into applications, and in the reverse direction, data integration in the lexicon from external sources must be correctly performed. The second reason is that, due to the fact that LMF addresses all languages and all NLP applications, the number of attributes is rather important, around 500. The methodology adopted was heavily influenced by the very nature of the object under study: NLP lexicons in a multilingual perspective. The solution to these two challenges has been to split the lexical specification into two separate objects: the structure and the content. LMF defines the structure while the features that encode information in form of attributes and its values are recorded in a Data Category Registry (e.g. Part-of-Speech). The advantage is two-fold: first, this registry, common to all TC37/SC4 standards, guarantees interoperability between e.g. lexicon and corpus annotation; and secondly, the peculiarities and requirements of different languages and linguistic schools are respected as recorded in the registry.

This is the essence of the “structure-adornment” binomial which neatly separates the standardization effort into high-level specification (the structure) and low-level specification (the adornment). The lexical information, the data categories to combine with the lexical model are even more crucial, since they allow implementation of the abstract model itself and development of standard-conformant lexical resources.

More precisely, LMF defines class names, class usages, class relations by means of English texts and UML diagrams. This specification goes with some guidelines and a series of examples, the so-called Lexical test-suites, i.e. practical examples associated to the international standards produced by the project to test the applicability and usability of the proposed concepts. These test suites accompany the LMF standard, facilitating its acceptance and implementation and promoting the development of LMF conformant lexicons. (Monachini 2007)

LMF is comprised of two types of packages:

1) The **core package** that consists of a structural skeleton in order to represent the basic hierarchy of information in a lexicon.
2) **Extensions to the core package** that reuse the core classes in conjunction with additional classes required for the description of the contents of a specific lexical resource.



The LMF model

The core package is specified by the following UML class model:

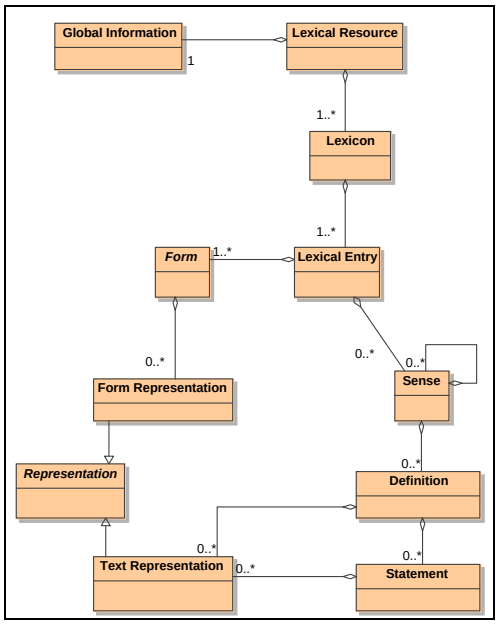


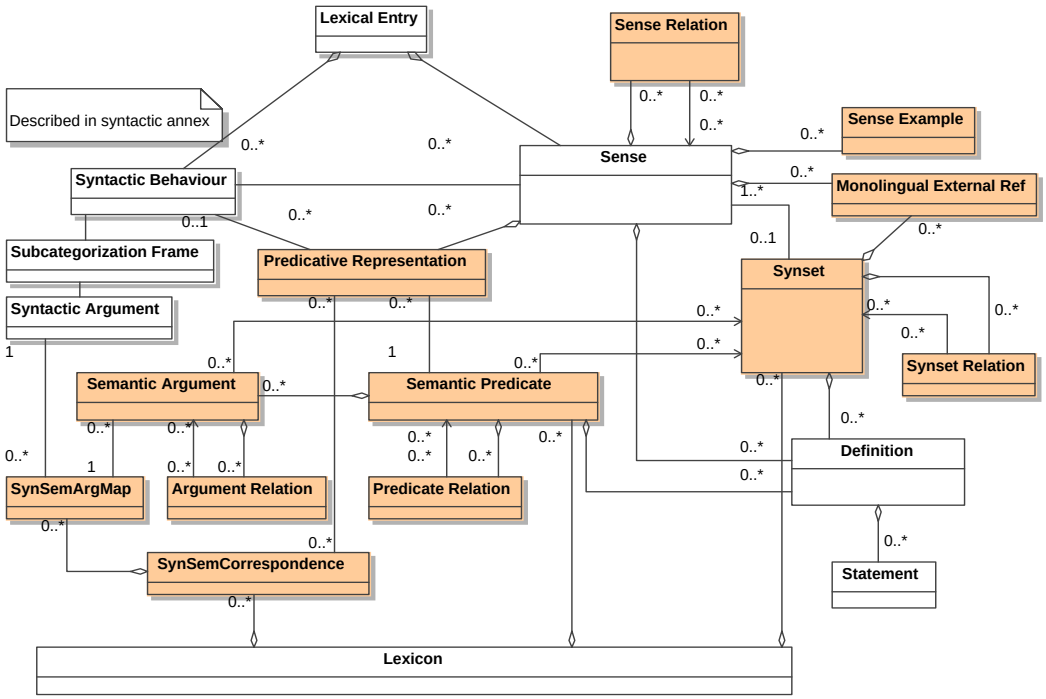
Figure 1. The LMF core model

The class called *Lexical Resource* represents the entire resource and is a container for one or more lexicons.
The *Global Information* class contains administrative information and other general attributes.
The *Lexicon* class is the container for all the lexical entries of the same language.
The *Lexical Entry* class is a container for managing the top level language instances.
The *Form* and *Sense* classes are parts of the *Lexical Entry*. Therefore, the *Lexical Entry* manages the relationship between sets of related forms and their senses.
(e.g. transliteration) the *Form* class may be associated with one to many *Form Representations*, if there is more than one orthography and one to many data categories describes the attributes of that orthography.

From the point of view of UML, an extension is a UML package. Current extensions for NLP dictionaries are: NLP Morphology, NLP Paradigm pattern, NLP Multiword expression pattern, NLP Syntax, NLP Semantics, Constraint Expression and Multilingual notations.

In the development of a lexical service ontology and implementation of services for lexicon access and interoperability, interoperability of semantic content is particularly crucial. For this reason, we concentrate here on the LMF extension which allows the representation of semantic information (figure below).

Semantic description departs from the core class *Sense*. The *Sense* class is associated with the *Lexical Entry* element and cannot be shared by two different entries. *Sense* node is the key element. It is not possible to describe *Synset* or *Predicate* instances without any *Sense* instance.
SynSet links synonymous sense instances. The LMF specification does not impose such strict guidelines on the exclusive usage of Sense and Synsets, i.e. they are not mutually exclusive.
Semantic Predicate describes an abstract meaning together with its association with the *Semantic Argument* class. In a lexicon, Predicate instances can be used to describe verbs and predicative nouns and Synsets for other meanings.
Semantic descriptions may be mapped to syntactic representations. More precisely, every *Semantic Argument* instance may be mapped to a *SyntacticArgument* of a subcategorization frame as defined in the LMF package for Syntax.



The LMF Semantic model

An XML DTD is based on the UML modeling, in order to allow instantiation of lexical entries conformant to the model.

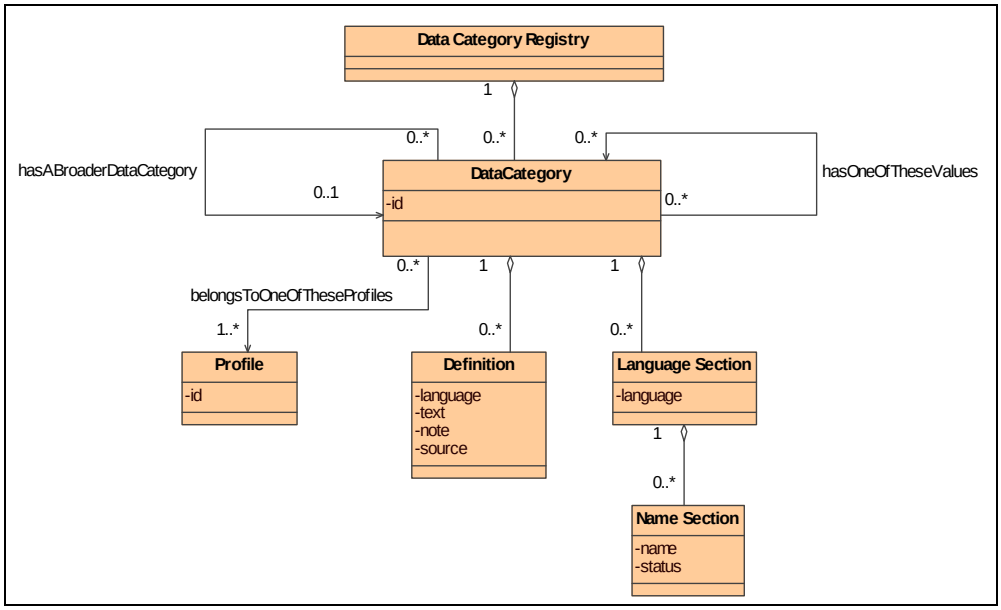
In order to allow the implementation of LMF constraints, the integration of LMF into semantic web applications, and, particularly the development of a lexical service ontology based on LMF, we defined an OWL format. The OWL Web Ontology Language is designed for use by applications that need to process the content of information. We use OWL in order to formalize a domain by defining classes and properties of those classes. OWL may be used to define individuals and assert properties about them, but we don't use these features. For some aspects, we use RDF statements that are basic object attribute value triples.
RDF/OWL specifications have been written from the UML model, according to the following rules: UML generalization is transcoded by means of "rdfs:subClassOf" tags. UML aggregation is transcoded as "owl:Restriction" and "owl:onProperty" tags. UML associations that are not aggregations are transcoded as "owl:DataTypeProperty" tags. Data category adornment is specified by means of a common super-class holding one or several "owl:Restriction" properties that are defined as "owl:ObjectProperty" with a pair of "rdfs:domain" and "rdfs:range".

3.3 Data Category Registry

The production of a family of consensual ISO specifications and data is critical for ensuring content interoperability and resource integration. The important task which is currently being conducted in parallel and in relation with LMF within ISO-TC37/SC4, is the work done in the Data Category Registry.

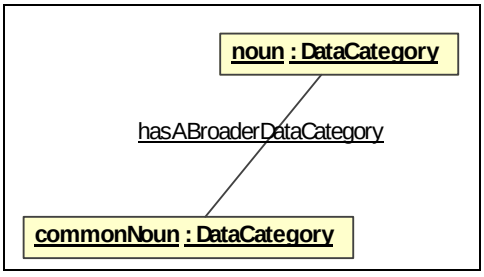
A Data Category is a *linguistic constant* representing the basic linguistic notions used to for the description of all languages. Data Categories provide the main building blocks of the lexicon, i.e. the descriptive components of the model, which practically make it possible to encode different lexical entries as instances of the abstract schema. Three sub-groups work in parallel (called 'Profiles'): one for morpho-syntax, one for syntax and one for semantics. Recently, a special Themantic Domain group dealing with lexico-semantic information has been set up (Project leader: M. Monachini; Chair(s): N. Calzolari and G. Francopoulo).

The current model DCR is based on Terminological Markup Framework (ISO-16642).



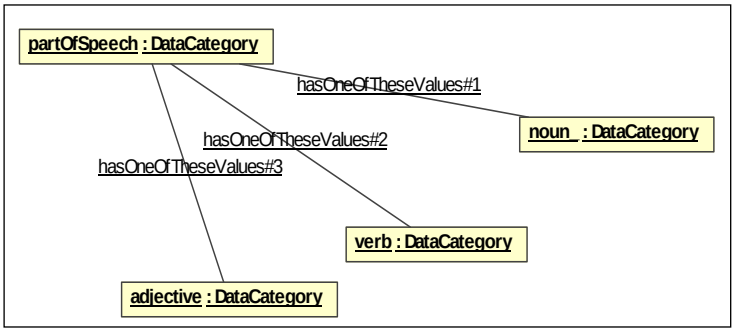
The DCR model

The notion of *broader relation* (in the figure below) allows us to define a hierarchy of constants, e.g. a common noun is a more specialized value than noun.



hasABroaderRelation

The notion of *conceptual domain* allows us to identify the set of valid values for a given attribute. As an example (see the figure below), in the Morphosyntactic profile, *noun*, *adjective*, *verb* are values allowed for partOfSpeech.



hasOneOfTheseValuesRelation

The current registry records values for West/East European languages and, to certain extent, for Semitic languages. Two other parallel tasks are currently being conducted. One task deals with Asian values within the NEDO project and the other one gathers values coming from the biomedical domain lexicon needed for the semantic representation of bio-terms.

3.4 An Ontology of Lexical Services

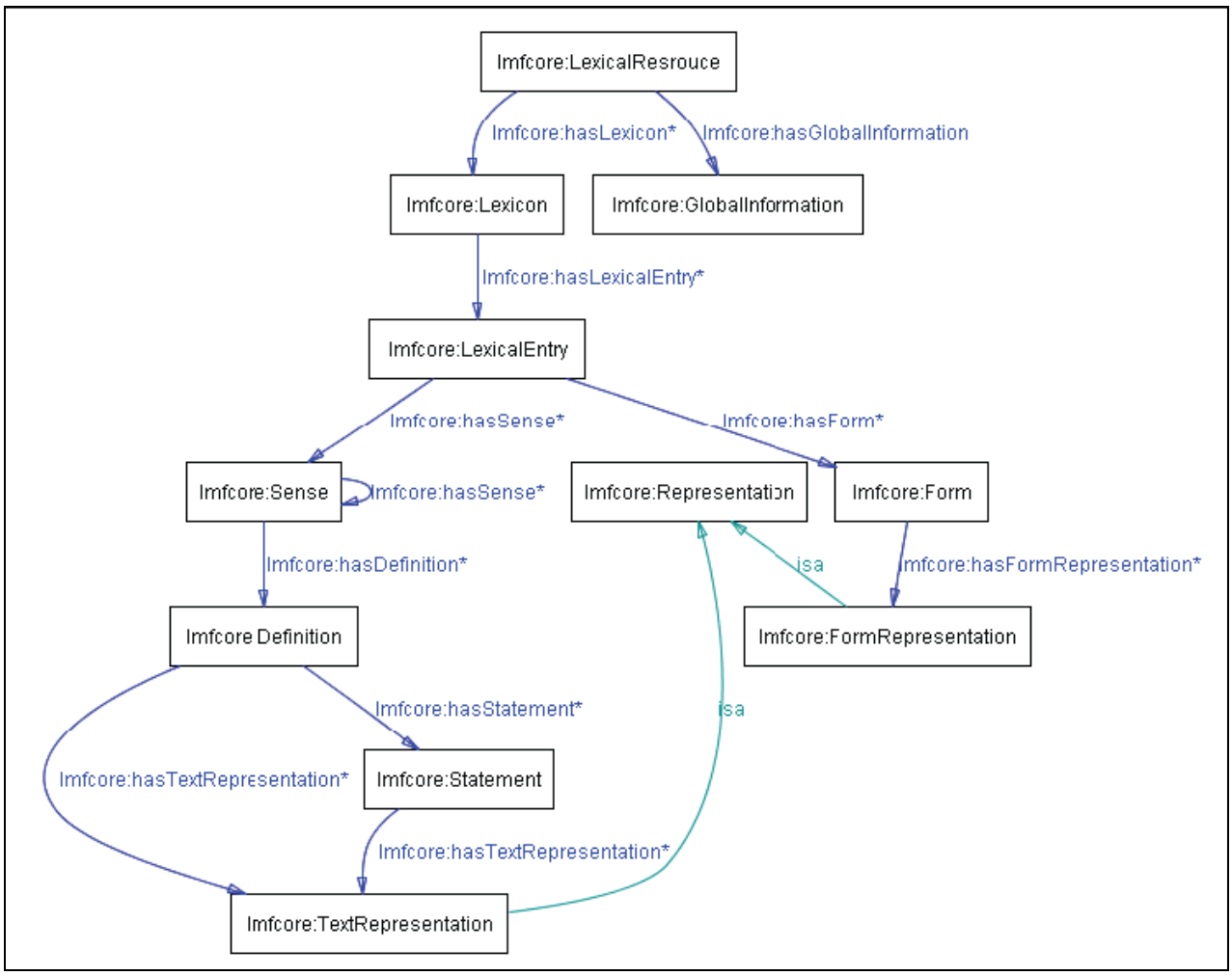
As already mentioned before, the provision of appropriate linguistic services in the heart of a global language infrastructure, implies the deployment of efficient workflows able to combine atomic services into composite ones to be consumed by end-user. The composition process calls for an ontology-based organization of both language resources and processing resources. In this environment, language resources need to be classified from a service-oriented perspective of a range of functions. Hence, to address the issue of interoperability, the taxonomy for resources should be ground to principled and shared ontology.

In cooperation between NICT and DFKI, an *ontologization*[5] of language resources employing the standardized modelling frameworks as defined in ISO TC37/SC4, has been provided.

As far as lexicons are concerned, in the language service ontology, the sub-ontology or taxonomy of lexicons defined based on a service-oriented perspective, may not be linguistically or lexicographically motivated. However, it would be far better to ground the service-oriented taxonomy to some lexicon ontology that is based on shared linguistic and lexicological principles. In order to implement the service- and ontology-oriented lexical side of the global infrastructure, we have employed LMF (Lexical Markup Framework) as standardized lexicon modelling framework, and utilized it as a foundation to stipulate the service-oriented lexicon taxonomy and the corresponding ontology for lexicon access functions[6]. Preliminary results of this activity have been reported in Hayashi et al, 2008a and 2008b.

As known, LMF, worked out by the ISO TC37/SC4 community, is in the final stage of the international standardization process. The specification of LMF (ISO24613, 2008) states that the ultimate goal of LMF is to create a modular structure that will facilitate true content interoperability across all aspects of electronic lexical resources. Given this goal, the proposed modular structure of LMF consists of a core package and a number of extensions for modelling a range of lexicons. These LMF extensions are presented by extending the LMF core package, encouraging us to ontologize them by organizing the classes defined in the core package as subclasses of the top LMF class. The figure below illustrates the ontological configuration for the LMF core model.

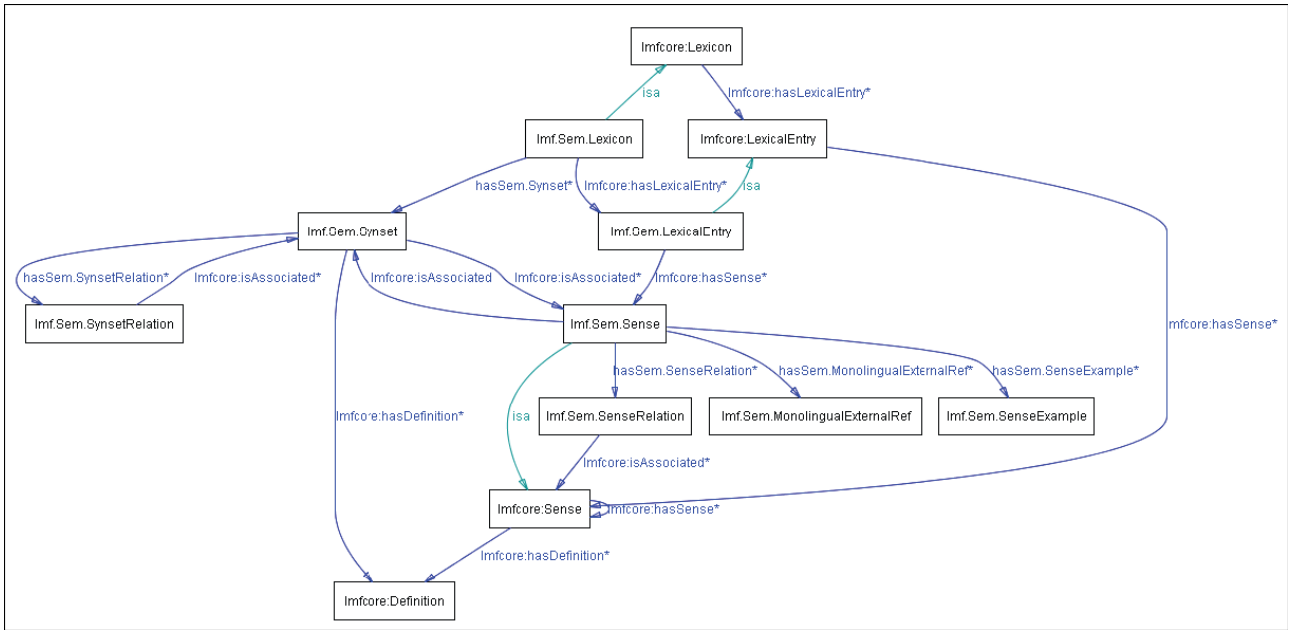
[5] Here "to ontologize" simply means to give a corresponding OWL representation to the constructs in the framework.
[6] On the side of services for linguistic annotation, the ontologization of Linguistic Annotation Framework (LAF), Morpho-Syntactic Annotation (MAF) and Syntactic Annotation (SynAF), has been provided.



Ontological Configuration of the LMF core model

The specifications of LMF are given by using UML (Unified Modelling Language) diagrams; these diagrams can be converted in a relatively straightforward way on OWL by applying some conversion conventions; for example, we have converted the aggregation in UML into hasXXX property. As stated in the figure, we have defined the LMF core package as an independent sub-ontology; the namespace lmfcore prefixed to the entities indicates this situation. All the other extensions are defined in another sub-ontology which imports the LMF core ontology; the namespace lmfall represents the whole sub-ontology. This account is somehow different from the original LMF specification, where, managing all types of lexicon in a single ontological space is not considered. However this account gives us an opportunity to stipulate a range of lexical resources in a unique ontological space, and this is mandatory in a service-oriented language infrastructure.

The figure below shows a part of the LMF NLP Semantics extension, which is associated in particular with the lexical semantic notions of the extension. This extension has been defined by sub-classing the classes in the LMF core package. The point is certain sub-class of the lexicon class is defined so as to have a particular type of the lexical entry. For example, lmf.Sem.Lexicon, as a sub-class of lmfcore:Lexicon, is defined as having lmf.Sem.LexicalEntry that is, in turn, a sub-class of lmfcore:LexicalEntry. Again, this account is somehow different from the original LMF specification, where, for example, sub-classing of the lexicon class is not allowed.



Ontological Configuration of the LMF semantic model

In an extremely simplified view of the lexicon taxonomy, the top-level class **`Lexicon`** includes a **`LexiconForNLP`**. This, on the other hand, derives a class for computational concept lexicon (**`ConceptLexicon`**), which has been introduced in order to stipulate WordNet-type lexical resources. The configuration of the service-oriented lexicon taxonomy can be quite arbitrary, rather than linguistically or lexicographically motivated. However, once we have ontologized the necessary parts of the LMF, we can ground the service-oriented lexicon taxonomy to the ontologized LMF. Each of the classes in the service-oriented taxonomy is defined in terms of lexical entry type that they accommodate, and the lexical entry types are defined in the ontologized LMF. When we have to represent and incorporate some new type of lexicon, we should first introduce a new lexical entry sub-class for the target lexicon in the service-oriented taxonomy, and then appropriately relate it to somewhere in the lexical entry taxonomy of the ontologized LMF.

3.5 Standardization Activities: The LMF-compliant NEDO Lexicon

While EAGLES and ISLE dealt with European languages, the Japanese NEDO project (Tokunaga et al, 2008), that develops international standards for Semantic Web applications, is specifically geared to Asian languages. NEDO contributed to ISO TC37/SC4 WG4 activities, by testing and ensuring the portability and applicability of LMF to the development of a description framework for NLP lexicons for Chinese, Japanese, Korean and Thai (and Italian). A major achievement has been the proposal of necessary extensions of the framework with respect to requirements and characteristics of Asian languages (in this is in line with the Multilanguage and multicultural mission of Language Grid). This activity culminated in the modeling of additional packages concerning the characteristics of Asian languages to be incorporated in LMF.

NEDO is developing a conceptual core for a multilingual ontology, with the main focus on Asian language diversity and a multilingual LMF-conformant core lexicon. Different from traditional approaches for designing a core lexicon, NEDO proposed a novel approach by starting from the Swadesh List of different language versions, such as English, Chinese, Bangla, Malay, Cantonese and Taiwanese. The list can be seen as a least common denominator for vocabulary. The coverage of the Swadesh list has been compared with the one of the Base Concept Set (BCS) as it is proposed by the Global WordNet Association. From a linguistic perspective, the NEDO lexical entries carry information coming from WordNet (synset membership) and, via cyclical mappings of English synsets' variants onto ItalWordNet (through the ILI), first, and onto the Italian SIMPLE semantic lexicon, afterwards, they have been integrated with further semantic information (e.g. linking to the SIMPLE ontology, semantic features, semantic relations, predicate argument structure). The preliminary experiment yielded promising results which motivate our ongoing work on other Asian languages. This core multilingual lexicon is being extended to be used and evaluated within a multilingual information retrieval system to access semantic content in a specific domain, sport and tourism,

in view of the Olympic Games. At the end, the NEDO lexicon will pose itself as an LMF-compliant lexical resource which the Lexicon Accessor service built on top of the GLI developed could be ready to access. A sample entry is provided below.

```
<LexicalEntry id="LE_fuoco_N"> <-- fire -->
        <feat att="POS" val="N"/>
        <Lemma>
                <feat att="writtenform" val="fuoco"/>
        </Lemma>
        <Sense id="USem60904fuoco" synset="N_1251">
                <feat att="semanticType" val="Phenomenon"/>
                <SenseRelation targets="USemD5364fenomeno">
                        <feat att="relation_type" val="lsa"/>
                </SenseRelation>
        </Sense>
<SyntacticBehaviour id="SB_SYNUfuocoN2 senses="USem60904fuoco" subcategorizationFrames="n-0"/>
</LexicalEntry>
<Synset id="N_1251">
        <feat att="ILI" val="N_8253345"/>
        <feat att="WN30_id_0" val="N_13480848"/>
        <feat att="WN30_weight_0" val="1"/>
</Synset>
```

A sample NEDO Lexical entry

3.6 Standardization Activities: An LMF compliant Special-domain Lexical Database

The lexicon described in this section should be considered as a customization of the LMF meta-model based on the requirements gathered from the biomedical community. The BioLexicon is a concrete example of application of the LMF framework. It poses itself as a standard for the representation of lexicons in the bio-domain, but thanks to its standard compliance, it could eventually be also interoperable with other lexicons in other domains.

The reasons behind the choice of the the ISO Lexical Markup Framework as the reference meta-model for the structure of the BioLexicon is that the biomedical field has strategic relevance today and research is being carried out to access its literature and extract knowledge. Access to and interoperability of biological databases, however, is still hampered by lack of uniformity and harmonisation of both formats and information encoded. A current demand in bioinformatics is to construct a comprehensive and incremental resource which integrates bio-terms encoded in existing different databases. A challenge is to encode all relevant properties of bio-terms according to the most accredited standards for the representation of lexical, terminological and conceptual information.

These assumptions are made operational in the design of the *BioLexicon* (Monachini et al, 2008). The BioLexicon is designed to be reusable and flexible in order to be used by different applications: e.g. information extraction and information retrieval. Since one of the main aims is to foster semantic interoperability in the biomedical community, the Lexical Markup Framework, together with the main building blocks for the representation of the entries used for lexical description – i.e. the Data Categories – provides a common, shared representation of lexical objects that allows for the encoding of rich linguistic information. The BioLexicon accounts for (English) terms related to the bio-domain and represent morphological, syntactic and lexical semantic properties of them. Among these terms, especially relevant here is the encoding of biologically relevant verbs and nominalized forms of verbs, i.e. verbs typically used in biomedical texts to refer to bio-events. For such lexical items a full explicit representation of their syntactic complementation and of their semantic argument structure will be represented. The BioLexicon thus encodes those linguistic pieces of information that domain ontologies partially lack and which are, instead, important for information and knowledge extraction purposes.

Another key property and research direction of extreme relevance for the Language Grid Infrastructure is that term entries in the BioLexicon will be linked to a BioOntology (a resource developed in parallel within the project) and both will serve as the terminological backbone for harvesting information from documents. A reusable BioLexicon with sophisticated linguistic information, linked to a bio-ontology, should enable the bio-informatic community to develop information extraction tools of higher quality.

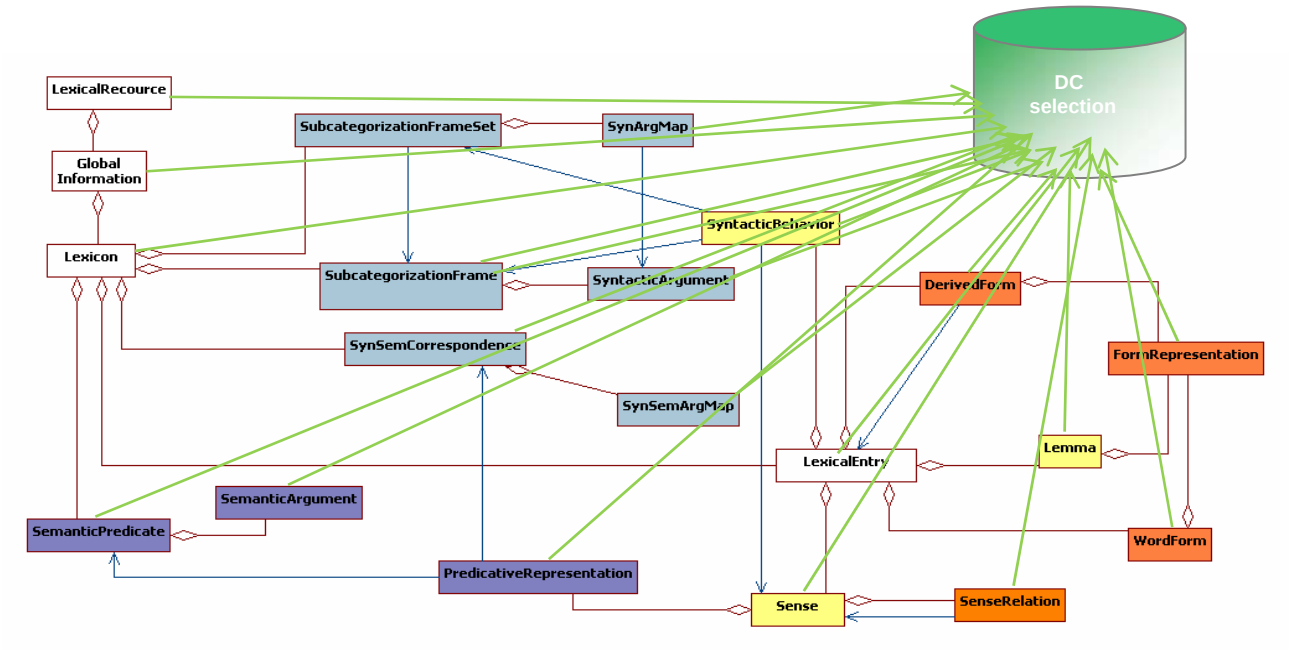
Another innovation of the BioLexicon is that the database comes equipped with software Java procedures for automatic uploading of the database. An *XML interchange format* (XIF) has been designed, on the basis of LMF, with the purpose of automatically populating the BioLexicon with data provided by domain experts and by lexical acquisition systems, therefore allowing for a standardisation of the data extracted from the different terminological resources and from texts (see section 4.1 below). This XML exchange format for the BL population, can facilitate the integration of the BioLexicon in the Infrastructure and be used as the input and/or output to the integrated tools either for updating new information into the lexicon or for retrieving lexical data (which are elements of a more composite lexical service).

3.6.1 BioLexicon Data Categories

Data Categories are the linguistic constants that are used to describe the single instances of the lexical classes. Data Categories take the form feature structures, or attribute-value pairs. In conformity to the ISO philosophy, the Data Category Selection for the BioLexicon is partially drawn from the ISO 12620 Data Category Registry (Francopoulo et al. 2008), and partially integrated by defining a set of specific DCs needed for the representation of the domain terminology, whenever missing from standard repositories. In order to be able to automatically constrain and check the consistency of the DCs on each specific object, in the BioLexicon, most DCs have been typed.

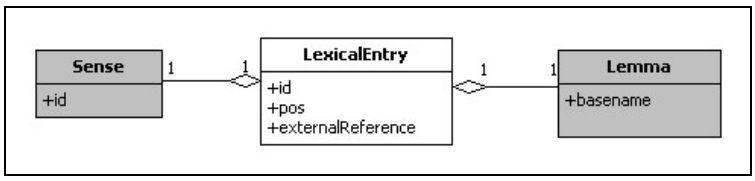
3.6.2 BioLexicon Model

The figure below offers an overall view of the BioLexicon lexical objects and of how they are decorated by the linguistic descriptors available in the repository of the BioLexicon Data Category Selection.



The BioLexicon data model and data category selection

The core lexical objects of the BioLexicon are: *LexicalEntry*, *Lemma*, and *Sense*. The *LexicalEntry* class represents the abstract units of vocabulary at three levels of description: morphology, syntax and semantics. To ensure modularity and extendibility the three levels of description are accounted for in separate lexical objects, independently linked to the LexicalEntry, which functions as a bridge among the *Lemma*, its related *Sense*(s), and *SyntacticBehavior*(s), which is the core unit of the syntactic layer. The *LexicalEntry* class represents the abstract units of vocabulary at three levels of description: morphology, syntax and semantics. To ensure modularity and extendibility the three levels of description are accounted for in separate lexical objects, independently linked to the LexicalEntry, which, thus, functions as a bridge among the Lemma – and their forms – its related Sense(s), and Syntactic Behavior(s). Lexical Entry bears a Part-Of-Speech DC, plus additional non mandatory attributes.

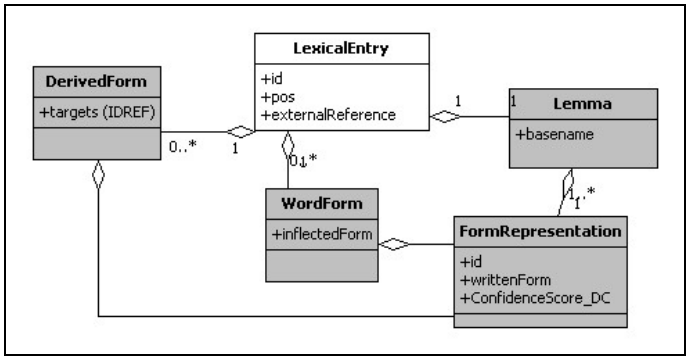


The BioLexicon core model

A specific requirement coming from the biology community is that the resource should keep track of the ids of the terms in other well known reference databases and ontology. External references in the BioLexicon are thus represented as typed data categories that are added as attributes to the *Lexical Entry* object. *Lemma* is used to represent the base form of lexemes plus additional grammatical properties; because it is in a one-to-one relation with the *Lexical Entry*, homonyms in the BioLexicon are represented as separate entries. Finally, the basic information units at the semantic level are senses. *Sense* is therefore the class used for the representation of the lexical meanings of a word/term, and it is inspired by the SIMPLE Semantic Unit (Ruimy et al. 2003). Each Sense instance represents and describes one meaning of a given Lexical Entry, contains information on the specific (sub)domain to which the sense applies, and contains a link to the Bio-ontology.

3.6.3 The morphological extension

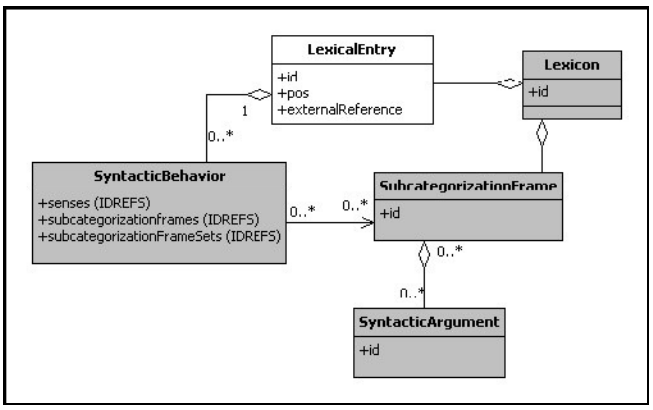
In a terminological lexicon for biology a key requirement is the representation of the different types of term variants. Given that linguistic information are automatically extracted from texts, in the BioLexcon we c distinguish only between two types of variants: variants of form and semantic variants. The morphology extension therefore has been implemented mainly to allow for a rich and extensible representation of variants of form. The *FormRepresentation* object has in fact the function of representing multiple orthographies. The basic DC specifying the *FormRepresentation* is the writtenform, i.e. the string identifying the form in question. Each variant is then adorned with properties represented by specific DCs: the type of variants ("orthographic", for variants and "preferred" for base forms), and a confidence score that the automatic extraction techniques assigned to each variant (for details on the treatment of variants see Quochi et al 2007). The *InflectedForm* class is used in the BioLexicon to represent the automatically generated inflected forms of domain-relevant verbs.



The morphological extension

3.6.4 The syntactic extension

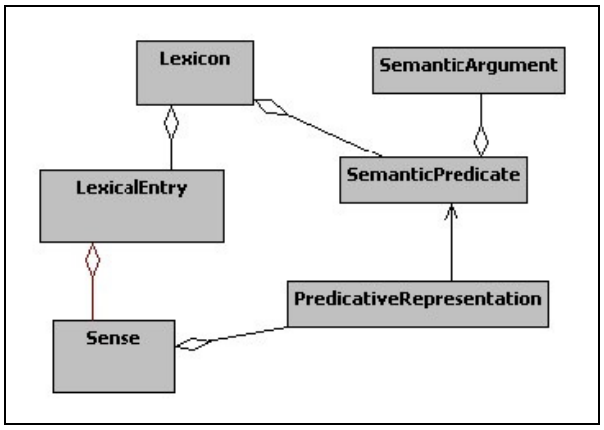
SyntacticBehavior represents one of the possible behaviors that a lexical entry shows in context. A detailed description of the syntactic behavior of a lexical entry is further defined by the *SubcategorisationFrame* object, which is the "hearth" of the syntax module. *Subcategorisation Frame* is used to represent one syntactic configuration and does not depend on individual syntactic units; rather it may be shared by different units. The LMF syntax extension is adapted in view of accommodating the subcategorisation behaviors of terminological verbs automatically extracted from texts by appropriate NLP algorithms, and thus a probability score will be recorded as a property of the Syntactic Behavior belonging to a give *SubcategorisationFrame*.



The syntactic extension

3.6.5 The semantic extension

The semantic module of the lexicon is made of lexical objects related to the Sense class. *Sense* represents lexical items as lexical semantic units. Semantic relatedness among terms is expressed through the *SenseRelation* class, which encodes (lexical) semantic relationships among instances of the Sense class. The BioLexicon *SemanticRelations* build on the 60 Extended Qualia relations of the SIMPLE model and are represented as Data Categories drawn from the Data Category Selection specifically defined to meet the needs of the bio-domain and of the BOOTStrep project (Monachini et al. 2007).
The *SemanticPredicate* class, instead, is independent from specific entries and represents an abstract meaning together with its associated semantic "arguments". It represents a meaning that may be shared by more senses that are not necessarily considered as synonyms. It is referred to by the *Predicative Representation* class, which represents the semantic behavior of lexical entries and senses in context, i.e. it describes the complete semantic argument structure of a predicative lexical item.



The semantic extension

4 Lexical Services on Global Language Infrastructure

In this section, we will describe some procedures developed at ILC in order to support the viability of the Language Grid service ontology and demonstrate some (composite) language services for accessing, querying, navigating LMF-compliant lexical databases and integrating them with other language resources. This is to be considered a step towards the standardization of lexicon access functions and the deployment of "LMF Lexical Web services".

4.1 SIMPLEtoLMF API

In order to export entries from the SIMPLE Italian lexicon into LMF, we have developed an API. This API allows performing queries to the lexical database from Java applications. This way, for each element of the entries to be exported, we apply queries from this API and encode it according to the LMF syntax.
As the database model is the same for the 12 lexicons (each for a different language) developed within the European project SIMPLE, the introduced procedures could be used out-of-the-box for SIMPLE lexicons for languages other than Italian.

4.2 "Ontologisation" of lexicons

An initiative carried out at ILC aimed at deploying lexicon access functions is the ontologization of SIMPLE, the lexico-semantic resource based on the Generative Lexicon (GL) as described in Toral and Monachini, 2007. This research aims at the representation of a lexicon based in the GL theory into the Semantic Web ontology language, with reasoning capabilities interfaced to a lexicon. The work consists in developing procedures to model the elements (i.e. semantic types, qualia relations, semantic features) from the SIMPLE original ontology into OWL[7]. A challenge in the ontology design and transformation is that its nodes are not only defined by their formal dimension (taxonomic hierarchy), but also by the GL qualia orthogonal dimensions: constitutive, telic, agentive[8].

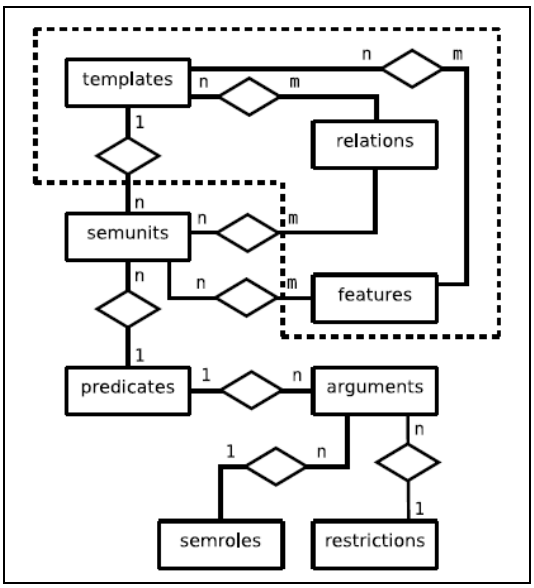
The ontologization of SIMPLE is not simply an automatic transformation process, but properly exemplifies typical lexical services, a kind of "communication protocol" aimed at interfacing/integrating a lexical repository with a conceptual one. Specific procedures have been developed that query the lexicon database and acquire specific semantic information which goes to enrich the OWL ontology. It should be mentioned here that both the lexicon and the (MySql) database architecture are fully compliant to LMF and can hence be exported in this standard format. This bottom-up approach, by exploring the word-senses belonging to a given semantic type and by using the qualia structure as a generative device, extracts from the lexicon features and selected constraints on relations that link semantic entries each other, thus promoting them at the level of semantic types. This allows this enriched and empowered ontology to be processed and checked by standard reasoners. This is useful for Semantic Web applications, semantic NLP tasks, and for enhancing the quality of the lexicon by validating it (through reasoning one can look for inconsistencies). The ontology is also a key element of a broader forthcoming research aimed at automatic lexico-semantic-driven text mining and knowledge acquisition procedures, which, in their turn, have the goal of gathering knowledge to enrich lexicon, thus creating a virtuous circle between lexicon/ontology and corpus-based information acquisition.

The figure below shows an entity-relationship diagram of the SIMPLE tables involved in these tasks. While only the three tables inside the dashed lines are employed for the transformation phase, all of them are used in the enrichment one.

4.2.1 Automatic transformation

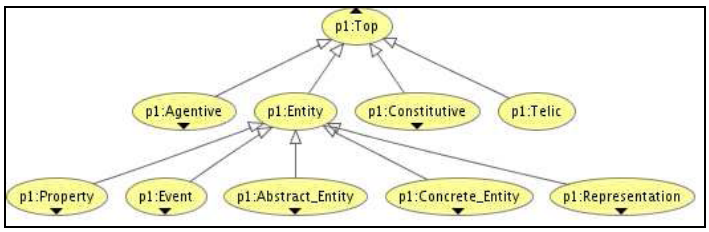
The transformation of the ontology involves translating the different elements that make up the ontology from their original codification as registers of database tables into OWL-compliant expressions. In order to carry out this task, we have written software that creates an OWL ontology by using the OWL API included in Jena (a Java framework for building Semantic Web applications). The input is provided by making queries to the original PSC database. From this database, ontology information was used for the transformation phase whereas mainly lexicon tables were queried for the enrichment step. Finally, in order to visualize and check the consistency of the created OWL ontology we have utilized the Protégé ontology editor with its OWL plug-in together with two OWL reasoners: FaCT++5 and Pellet.

[7] For an OWL version of WordNet, cf. Van Assem 2006.
[8] Constitutive: composition of entities; Telic: function of entities; Agentive: origin of entities.



Tables of the SIMPLE data base involved

Four different elements of the ontology have been identified and transformed by means of an automatic procedure. These are the ontology taxonomy, relations, features and cardinality restrictions that apply to ontology nodes on both relations and features. A detailed explanation about the translation into OWL of each of these ontology elements follows. The taxonomy of classes is derived from the templates table. First, all the different semantic types are identified and the correspondent OWL classes are created. Next, the taxonomy is built by identifying for each class its direct ascendant and making the latter explicitly the super class of the first. Finally, all siblings across the class taxonomy are made disjoint.



The SIMPLE OWL Top Taxonomy

Relations are extracted from the relations table. As in the case of the templates, a taxonomy has been built for relations. The top nodes are the different relation types present in the SIMPLE model, i.e. four types for the correspondent qualia roles (agentive, constitutive, formal and telic) and others for non-qualia relations (antonym, derivational, metaphor, metonymy, polysemy and synonym). Domain and range are both set to the top node of the ontology for non-qualia relations while for qualia relations both are set to the ontology classes "Entity" and the class that corresponds to the specific qualia type ("Agentive", "Constitutive", "Formal" or "Telic").

Features are imported from the features and templates tables. Differently than for relations, features form a plain taxonomy, i.e. there are not sub-properties relationships. Templates information is used to establish the domain, as this is defined as the union of classes for which the feature is defined. The range is set to Boolean as so far only these kinds of features have been imported. Finally, cardinality restrictions are imported from the three tables depicted in figure above inside the dashed lines. They are extracted in a top-down fashion so that the procedure can deal with inheritance. For each class we first check if the current restriction is already inherited from a super-class. Only if it is not, then the restriction is applied. The procedure has found 13 inconsistencies in the ontology database, i.e. restrictions inherited and explicitly encoded with different cardinality values. Thus, the procedure has been useful also to check and improve the input resource.

4.2.2 Enrichment

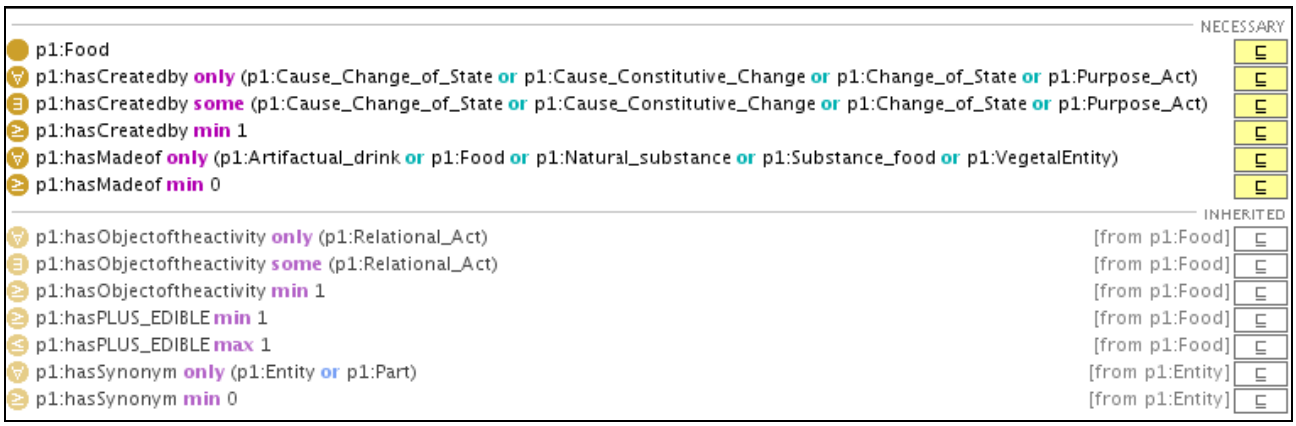
The enrichment phase extracts from the lexicon further information not present in the original SIMPLW ontology and, in most of the cases, automatically adds it to the OWL ontology. Different kinds of knowledge are extracted this way: quantifier restrictions, predicates and additional features and relations. Within ontology, quantifier restrictions allow to establish, for a restriction applied over a property to a source class,

the target class/es of this restriction. There are two different types of quantifiers: existential and universal[9]. Despite of the fact that the SIMPLE ontology does not contain the semantic types that are the target to a given restriction, this information can indirectly been extracted from the lexicon and, after some generalization, be used to enrich the ontology. For a given constraint over a relation that belongs to a template, we extract all the occurrences of the relation in the semantic units that belong to the template's semantic type. These are made up of a source semantic unit that belongs to the current semantic type and a target semantic unit. I.e. they link two semantic units. E.g. the semantic unit "bisturi (scalpel)" that belongs to the semantic type "Instrument" is linked by the relation "usedBy" to the semantic unit "chirurgo (surgeon)" that belongs to the semantic type "Profession". For each of these occurrences, we extract the semantic type to which the target semantic unit of the relation belongs. Therefore, we obtain a list of target semantic types. Afterwards, these are generalized in this way: if in the list it is present a semantic type and one ancestor of it, then the descendant semantic type is deleted from the list. For example, there are 47 semantic units in the semantic type "Food" that instantiate the telic relation "Objectoftheactivity", out of which we obtain the target class "Relational Act". Regarding the quantifier type, we add an universal restriction to all the constraints while existential restrictions are only applied to that constraints of type "Yes" or "RecYes"[10] as an existential restriction implies a minimum cardinality greater than zero. Following with the previous example, both an existential and universal quantifier restrictions would be added for the relation "Objectoftheactivity" as its constraint value is "RecYes".

Although semantic predicates are not included in SIMPLE at the ontology level, they are defined in the lexicon. The challenge consists then in establishing generic predicates for the nodes of the ontology of a predicative nature (the "Events" semantic type and its subclasses) by generalizing them from the predicates present for the semantic units that belong to these semantic types. Concretely, given a semantic type and a set of predicates (those of the corresponding semantic units), we generalize the selectional restrictions that belong to each of the different predicative semantic roles to one or more semantic types. A clear parallelism can be established between this issue and that introduced above as also here we have to generalize the target of relationships to semantic types. The difference, however, is that the previous case consisted in finding for a set of semantic units the corresponding semantic types whereas in this case not only semantic units need to be translated into semantic types but also notions (a selectional restriction can be a semantic type, semantic unit or a notion). Afterwards, a quantifier restriction is introduced over each predicative semantic role relation. The target of the restriction is the semantic type/s result of the generalization of the gathered set of semantic types. Regarding the enrichment phase, through the automatic procedures developed we obtain a language independent enriched ontology from language-dependent (Italian) lexico-semantic information. The figure below shows the asserted conditions present in the node "Artifact Food" of the output ontology. Two different areas are presented; the upper one includes the necessary conditions, those specific of the class, whereas in the lower part we find the inherited conditions, those that the current class takes from its super classes by inheritance. For each relation we can see in the resulting ontology the corresponding cardinality and quantifier restrictions, the latter including target classes extracted from the lexicon. Regarding the only feature present in the original template, "Plus Edible", the correspondent minimum and maximum cardinality restrictions are shown in the inherited part of the figure as the direct super class ("Food") introduces as well these constraints and thus there is no need to explicitly repeat the same information for the class "Artifact Food".

[9] An existential restriction describes the set of individuals that, for a given property, have at least one relationship with individuals that are members of the target class. On the other hand, an universal restriction describes the set of individuals that, for a given property, only have relationships with individuals that are members of the target class.

[10] i.e. if the semantic relation is considered obligatory in the Template table of the SIMPLE database.



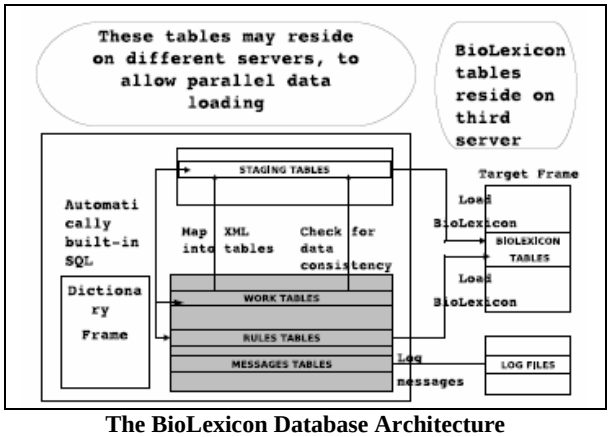
Enrichment of Food type with constraints on relations between the type itself and other semantic types

The result of the current research is a semantically rich ontology with reasoning capabilities interfaced to a lexicon. Therefore it is a valuable resource for semantic Natural Language Processing tasks. In fact, this ontology is a key element of a broader future research which is aimed at guiding automatic lexico-semantic Text Mining and Knowledge Acquisition procedures.

4.3 Automatic Population of the BioLexicon

The BioLexicon conceptual model has been implemented as a relational database capable of managing biological data both extracted from texts and collected from other existent resources. The BioLexicon DataBase (BLDB) consists of two modules: a MySQL database, and a java software component for the automatic population of the database. External to the BLDB, but fundamental for its automatic population, is an XML interchange format (XIF), which the java procedures parse and read to load data into the BLDB. The XIF thus allows for a standardization of the data extracted from the different terminological resources and from texts (by automatic NLP applications) and for the independency of both the uploading procedures and the BLDB from native data formats. The database is structured into three logically distinct layers:

1. the DICTIONARY FRAME contains tables used in the first handling of the XML Interchange Format and it rules that automatically build SQL instructions to populate target tables;
2. the STAGING FRAME is set of hybrid tables for volatile data;
3. the TARGET FRAME contains the actual BioLexicon tables i.e. tables that directly instantiate the BioLexicon DTD and contain the final data.



The BioLexicon Database Architecture

The neat separation between target tables (the BioLexicon proper) and "operational" tables allows for the optimization of the data uploading into the BLDB and ensures an easy extendibility both of the database and of the uploading procedures. In the near future, the database will be integrated in a UIMA framework and accessed either through APIs by software users or through a web graphic interface by various types of users with different needs. At present, the BLDB can be accessed and queried locally through a prototype graphic interface. Currently, the BLDB contains terms and variants gathered by existing resources, with derived relations, and a set of automatically generated verb forms.

Here we describe the three levels of the database "at work". As explained before, input data is not directly loaded from the original resource, but is loaded from the XIF. As shown in the XIF fragment below the Cluster element contains a set of coherent data encoded in specific sub-elements that represent linguistic notions. The Extraction Transformation Loading (ETL) process extracts (E) data from the input files, transforms (T) and loads it in temporal tables (staging tables) and finally loads (L) it in the actual database tables (the target tables). The dictionary level of BLDB is logically divided into two separated parts: WORK and RULE. The former manages the mapping of the XIF onto staging tables (E-T phases), while the latter deals with the upload of data into target tables (L-phase). Staging tables have been modeled to be in a one-to-one correspondence with the XIF elements. Clearly, also the element attributes are mapped to staging columns. Let us consider the following example:

```
<Cluster clsId="SC494014" SEMTYPE="GeneProt">
<Entry entryId="SC494014_1"
baseForm="Isopullulanase precursor"
type="PREFERRED">
<SOURCEDC sourceName="UniProt"
sourceid="O00098"/>
<POSDC posname="POS" pos="N"></POSDC>
<Variant writtenForm="isopullulanase gene"
type="orthographic"/>
<DC att="swissprot_name" val="CISY_EMENI"/>
<DC att="speciesNameNCBI" val="162425"/>
</Entry>
</Cluster>
```

The WORK part of the Dictionary (WORK henceforth) maps XIF elements onto staging tables. Due to the design of the conceptual model, we decided to implement relations among objects as correspondence tables. For instance, the Variant element in the XIF determines also the correspondence table between Lemma and FormRepresentation tables. This means that, while FormRepresentation contains a list of variants, the Lemma_FormRepresentation table contains variants defined for a given lemma. This is crucial since in the biological domain, the same orthographic form can be a variant of different lemmas.
Correspondence tables are defined both at staging and target level. Staging tables, therefore, contain raw data, which has to be subsequently manipulated in order to be loaded into target tables. Let us consider, for instance, how the Variant element instantiates the FormRepresentation and the Lemma_FormRepresentation staging tables.

```
<Entry entryId="SC494014_1"
baseForm="Isopullulanase precursor"
type="PREFERRED">
<Variant writtenForm="isopullulanase
gene" type="orthographic"/>
</Entry>
```

WORK encodes information about the Entry and Variant elements. In details, it "knows" that the Variant element has its own identifier and that this identifier is built with a fixed rule. WORK also "knows" that the same element defines a correspondence table between itself and its parent element (Entry). Let us show below how WORK creates input files for staging tables (for FormRepresentation and Lemma_FormRepresentation respectively):

```
"FR_isopullulanase gene", "isopullulanase gene", "orthographic"
"LM_Isopullulanase precursor","isopullulanase gene"
```

The direct benefit of using the dictionary level is that the loading software builds "objects" on the basis of XIF elements contained at dictionary level and manages only these objects. This means that the mapping between XIF and staging tables is performed only once, during the E-T phase. Even the I/O operations are performed once per object as well as the loading of the data in the tables. The second part of the dictionary is the RULE one (RULE hereafter). This part manages the mapping between staging and target tables and regulates the L-phase. This mapping is required since there is no one-to-one mapping between staging and target tables. RULE, therefore, maps source staging tables onto target tables and allows for the automatic creation of SQL instructions. These instructions are simply "SELECT..FROM...WHERE..." that, when executed, retrieve data from staging tables and save them in input files for target tables. We adopted this

strategy to allow wide freedom in defining rules to populate target tables. A typical example of L-phases is the decoding process that leads from the attribute-value pair to the corresponding identifier. Data categories, for example, are encoded in tables that are managed at L-phase of the loading process. The staging FormRepresentation table contains the value "orthographic", which identifies a type of variant. A data category VariantDC decodes this value in an identifier. RULE creates the following SQL instruction:

```
"SELECT a.id,a.writtenform,d.id
FROM FormrepResentation a, VariantDC b
WHERE a.type=b.val".
```

When executed, this instruction produces the following input file ready to be loaded in the target FormRepresentation table:

```
"FR_ isopullulanase gene", "isopullulanase gene",
```

RULE also creates objects on the target tables, which manage at once input files, SQL instructions and other features. In conclusion, we can see the dictionary level as a middleware between the original data, encoded in XIF, and the actual database. This structure of the database allows speeding up the loading process, since it is split into two different phases,: i) from XIF to staging tables and ii) from staging to target tables. Just to add statistical information, all chemical data (more than 100,000 entries) are loaded in less than two minutes.

4.3.1 Integration of the BioLexicon in Infrastructure(s)

The XML exchange format derived from LMF and devised for the automatic BL population procedure can be useful in view of the integration of the BioLexicon in an infrastructure based on LMF language services. This exchange format could be used as the input and/or output to a pipeline of integrated tools either for updating new information into the lexicon or for retrieving lexical data. One of the potential functionalities of the format could be also to merge information extracted by different tools, provided they use the same IDs and tagsets. Some Java procedures for the uploading of lexical data have developed on the basis of this XML format. This ensures compliance and reusability of the software for different application domains. The database behaves as a Common Analisys Structures consumer and a Common Analisys Structures producer, guaranteeing that the output is more detailed than the input.

4.4 Unifying Lexica and Composing Services "on-demand"

We show here some steps towards the concrete realization of the largely discussed and invoked paradigm of interoperable lexicons.

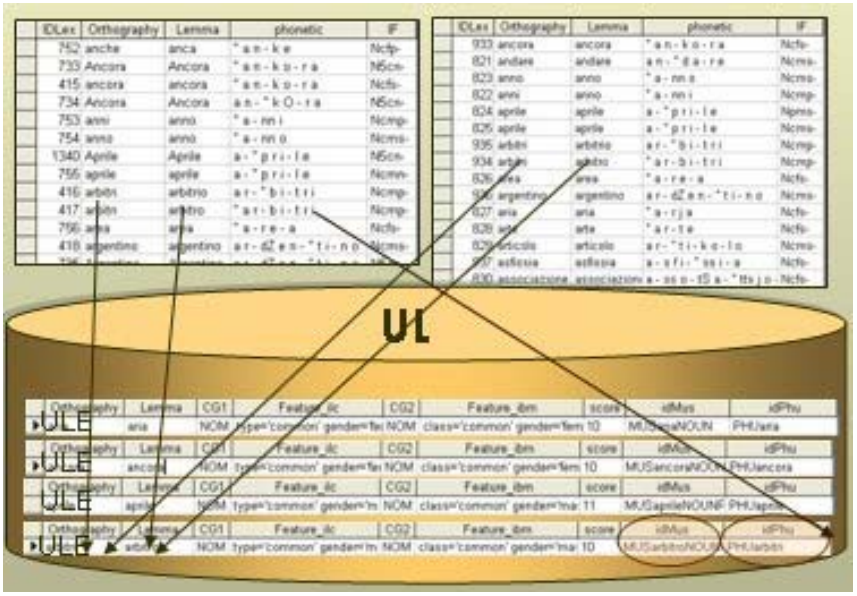
On the one hand, a procedure for merging different monolingual lexicons is described (Monachini et al, 2006). The assumption behind unification/merging/integration activities is that development, packaging and customization of LRs are critical in view of stimulating the industrial market (but also academic institutions). As concerns lexicons, the market is increasingly calling for new types of resources that can be built rapidly – tailored to specific requirements – possibly by combining certain types of information from available lexicons while discarding others. The ideal resource to fulfill users' requirements is often difficult to find. Conversely, the LR landscape can offer a large number of individual resources that could potentially contain what meets users' expectations. The problem is that, since they have been created by different developers for different purposes, these resources cover different types of data and linguistic phenomena; and, what's more, the information can be expressed in diverging formats. In this scenario, mapping and merging of resources yield a practicable and viable solution to make the available material usable, while protocols/services in this direction would boost its effective exploitation.

In Bertagna et al, 2007, we move from the hypothesis that having lexicons as distributed lexical repositories available via web services would allow creating new resources on the basis of existing ones, to exchange and integrate information across repositories and to compose new services on demand. This new type of LRs can still be stored locally, but their maintenance and exploitation can be a matter of agents choreographed to act over them. In the development of an infrastructure in the form of distributed language services, multilingual issues are of foremost importance. Large-scale multilingual lexicons are not yet as widely available as

needed, though they are the cornerstone of several multilingual applications. A new trend tries to exploit the richness of existing lexicons, in addition to creating new ones[11].

4.4.1 Unified Lexicon

An initiative in this direction, the *Unified Lexicon* project, has been carried out jointly at ELRA by its Production Committee and ILC-CNR (Monachini et al, 2006). The first goal of this work was to illustrate a specific procedure for merging different monolingual lexicons, focusing on techniques for detecting and mapping equivalent lexical entries. The experiment consisted in linking the LC-Star and PAROLE lexicons to set up a methodology to connect Spoken and Written LRs and obtain, as a by-product of this unification, Unified Morphosyntactic Lexicon Specifications (in line with the ISO directives).



The mapping and unification procedure

However, the major contribution of this experiment resides not solely in the implementation of the unification procedure *per se*, but, in the definition of an effective lexicon production model, which is attractive for producers and users. The model has been called "Unified Lexica on demand", a new paradigm of LR construction, which enables the community to customize available individual language resources via unification. In the envisaged scenario, the same lexicons may be made available to different users, who can select different portions of the same lexicon or combine information coming from different lexicons.
This is particularly crucial when considering the principles and requisites underlying a Lexical Grid: resources should be handled and procedures executed "on demand" on the basis of both user requirements and internal workflow rules. The Lexical Grid should help overcoming the notion of static and closed resources by offering appropriate services as a response to the need of merged and/or combined lexica and procedures that make resources openly customizable.

4.4.2 LeXFlow: An Architecture for Merging Lexical Resources

To meet the needs of a distributed language and service infrastructure, we have designed and built an architecture, LeXFlow, enabling a rapid prototyping of cooperative applications for integrating lexical resources (Soria et al, 2006)[12]. It is a web application, based on a web-service architecture, fostering integration, cooperative and collective creation and management of computational lexicons, addressing semi-automatic integration of computational lexicons, with focus on linking and cross-lingual enrichment of distributed LRs. As case–studies, we have chosen to work with:

[11] Admittedly, this is a long-term scenario requiring the contribution of many actors and initiatives particularly, international cooperation. In this sense, Language Grid is doing steps forward to realize this vision.

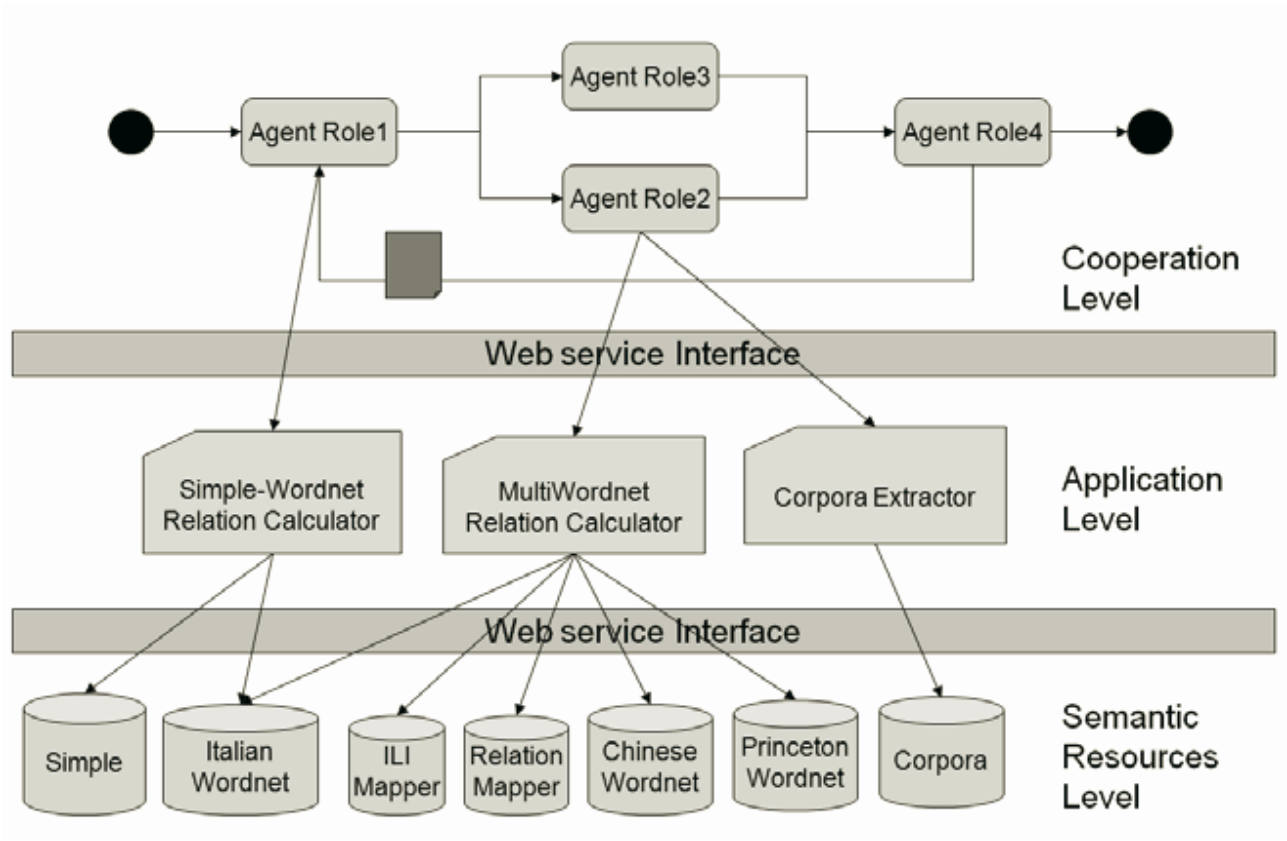
[12] LeXFlow has been recently integrated with the SemKey module (Marchetti et al., 2007), especially dedicated to collaborative semantic annotation of documents and textual excerpts, also referred to as social semantic tagging.

i) two Italian lexicons based on different models, SIMPLE and ItalWordNet, and
ii) two lexicons belonging to the WordNet family, ItalWordNet and the Chinese Sinica BOW.

These represent different opportunities of adopting a bottom–up approach to exploring interoperability for lexicon augmentation and mutual enrichment of lexical resources, either i) in a cross-model or ii) in a cross-lingual enrichment/ fertilization of monolingual lexicons (cf. figure below).

In the case of integration of lexicons with different underlying linguistic models, the availability of MILE (Calzolari et al, 2003), now of LMF, is an essential prerequisite[13].

From a more general viewpoint, we must note that the realization of the new vision of distributed and interoperable LRs is strictly intertwined with at least two prerequisites. On the one side, LRs need to be available over the web; on the other, the LR community will have to reconsider current distribution policies, and investigate the possibility of developing an "Open Source" concept for LRs.



LeXFlow Three-layered architecture

Multilingual WordNet Service

This module is responsible for the automatic cross-lingual fertilisation of lexicons with a wordnet-like structure. Put it very simply, the idea behind it is that a monolingual WordNet can be enriched by accessing the semantic information encoded in corresponding entries of other monolingual WordNets. The various WordNet-lexicons reside over distributed servers and can be queried through web service interfaces. The entire mechanism is based on the exploitation of the Interlingual Index (ILI). The proposal to make distributed WordNets interoperable allows applications such as:

- Enriching existing resources. Information is not complete in any WordNet: by making WordNets interoperable we can bootstrap semantic relations and other information from other WordNets.

[13] In LeXFlow, the MWNS presupposes the shared and conventionalized architecture of the WordNet framework. Our system is able to rely on it without resorting to the more comprehensive standard LMF.

- Creation of new resources. Multilingual lexicons can be bootstrapped by linking different language WordNets through the ILI.

- Validation of resources. Semantic relations and synset assignments can be validated if reinforced by data coming from other WordNets.

This work can be a prototype of a web application to support the Global WordNet Grid initiative (www.globalwordnet.org/) (Fellbaum and Vossen, 2007), whose success depends on whether there will be tools to access and manipulate the rich internal semantic structure of distributed multilingual WordNets. LeXFlow offers such a tool, providing interoperable web-services to access distributed WordNets on the grid. This allows exploiting in a cross-lingual framework the wealth of monolingual lexical information built in the last decade. As an example of use, a multilingual query given in Italian but intended for querying English, Chinese, French, German, and Czech texts, can be sent to 5 different nodes on the Grid for query expansion, as well as performing the query itself. This way, language-specific query techniques can be applied in parallel to achieve results that can be then integrated. As multilingualism clearly becomes one of the major challenges of the future of web-based knowledge engineering, WordNet emerges as a leading candidate for a shared platform, representing a simple and clear lexical knowledge model for different languages. This is true even if it has to be recognized that the WordNet model is lacking some important semantic information (like a way to represent semantic predicates).

4.5 UFRA: A UIMA-based Approach to Federated Language Resource Architecture

Integration of both LRs (lexicons, ontologies, corpora, etc.) and various NLP tools into a common framework of shared and distributed resources is being pursued at ILC by using the IBM UIMA middleware (Ferrucci and Lally, 2004). In Del Gratta et al, 2008a and Del Gratta et al, 2008b various research initiatives, experiments and case studies towards achieving interoperability and integration between a lexical resource and text annotation in the framework of a "UIMA-based" platform are reported. Background reasons are that UIMA provides a useful middleware for integrating linguistic services according to a language service ontology and to user needs. The first prototype is developed along the lines provided by the Language Grid project, since it inherits its service ontology environment.

In our approach, called UFRA (Del Gratta et al, 2008b), however, we extend the Federate Database Architecture System (FDBS) adding typical functionalities coming from UIMA. This approach is preferred to a standard resource-sharing architecture, since the FDBS also manages users and roles definition.

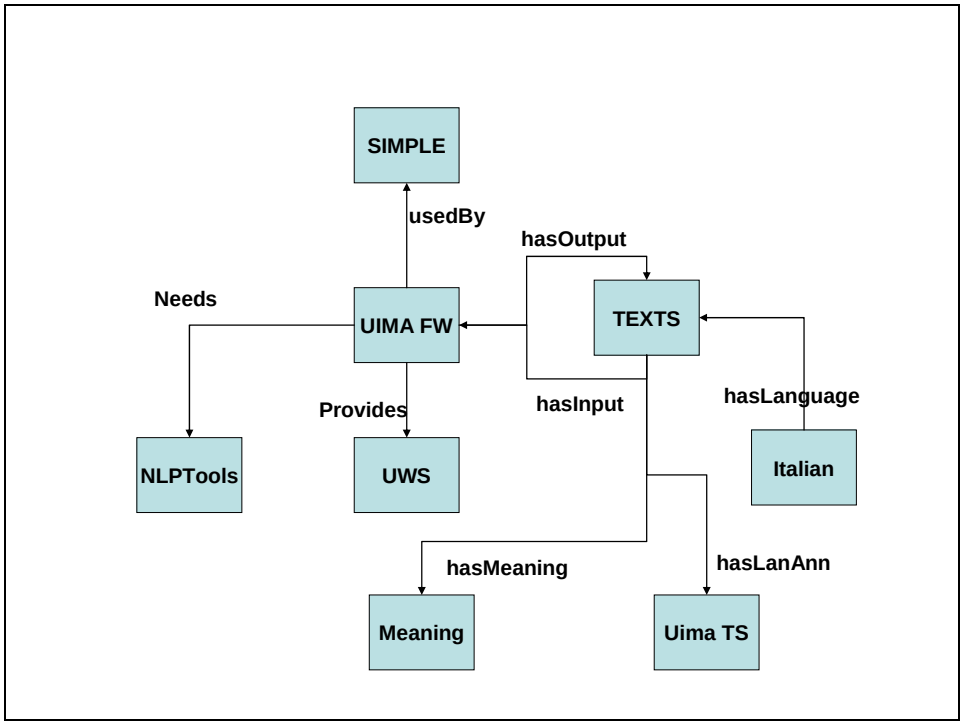
Here, the idea is that the user provides a text in input and selects the annotation (for example syntactic functional annotation); the platform then should be able to build an adequate pipeline that will provide the requested output. The language service ontology is the fundamental resource that enables to build pipelines: it cooperates with the list of available services (stored in a registry). The ontological properties of every single service keep track of the relations existing between such service and the others services offered by the platform. For instance, in order to obtain a functional annotation, we need intermediate analysis steps such as tokenization and lemmatization. So, on the one hand we have a user who asks for a final service among those available and, on the other, we have the platform, which internally auto-organizes in order to provide that service. The Language Grid comes into play in the definition of the service ontology.

The FDBS clearly defines a central authority responsible for all the interoperability outcomes among components and for standardization of input/output formats as well as resource structure. This central authority oversees to the federation policies; internal rules, groups and user rights and component cooperation are, indeed, the pillars of FDBS architecture. Fundamental in such architecture is a resource registry. In this way, we capitalize on the advantages of a federated architecture, such as autonomy, heterogeneity and distribution of components, monitored by a central authority responsible for checking both the integration of components and user rights on performing different tasks. We use the UIMA approach to manage and define one common front-end interface, enabling users and clients to query, retrieve and use language resources and technologies. In UFRA, we adopted the CLARIN strategy with respect to the setting up of the resource registry. Such a repository, defined following a standard metadata set (Broeder and Wittenburg. 2006), is the backbone of the UFRA architecture: it is used for both resource querying and services providing. The repository defined above is accessed, internally, by the UIMA framework and

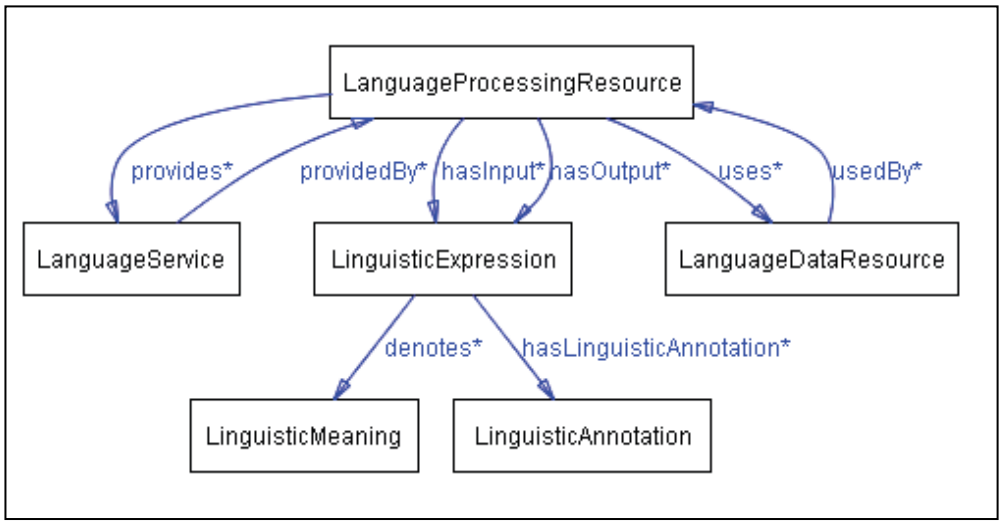
externally, by users who want to build their own resource collection relevant for their research. To identify one single resource with a primitive analysis engine is relatively straightforward: these analysis engines can be easily deployed as web services.

An activity related to Language Grid, is the use of UIMA to create a service for accessing the SIMPLE semantic lexicon within the task of annotation of temporal relations (Del Gratta et al, 2008a), by using the TimeML annotation schema (Pustejovsky et al, 2003). This annotation is integrated within the Italian Treebank and the SIMPLE lexicon. This research intends to contribute both to a UIMA type systems standardisation and to the definition of a common framework for resource and tool sharing and interoperability.

The figures below compare, on the one hand, the work-flow of components needed for the task of TimeML annotation and, on the other hand, the top level Language Service Ontology.



Work-flow of Components for temporal annotation



Top Level Language Service Ontology

We can see that the two figures are quite similar, even if the LanguageProcessingResource, as shaped in the second figure, is transformed in a more complex system in the previous one. In the SIMPLE-TimeML integration, the LanguageProcessingResource is a combination of the UIMA framework, heuristics, and a set of NLP tools especially defined to help heuristics working properly. For the sake of clarity, the integration of SIMPLE and TimeML allows to signal potential event denoting words in a text, simply taking decisions on the semantic types of SIMPLE and on the implementation of heuristics to detect events.

The workflow in fig2, therefore, can be rearranged by defining the following logical equivalences:

- LanguageProcessingResource (LPR): UIMA Framework and NLP tools, such as sentence and token annotators

- LanguageDataResource (LDR): the SIMPLE database.

- LinguisticExpression (LE): The text, with its language and (when available, a standoff annotation).

- LanguageService (LS): Uima annotators can be deployed as web services and remotely executed.

One service is built upon a set of linguistic features, i.e. of linguistic annotations.

- LinguisticMeaning(LM): the sense of the word, as encoded in the lexical database.

- LinguisticAnnotation(LA): an annotation. This is carried out by the UIMA type systems.

So the top level Language Service Ontology, when projected in the actual SIMPLE-TimeML mapping, can be interpreted as follows:

LRP uses the LDR to create lists of events-denoting words. LRP is also usedBy the LDR as feedback. The NLP tools contain rules that have to be matched against the LDR. NLP tools and the SIMPLE lexicon then cooperate to define the list of words.

LPR hasInput a LE, since it reads from files. LPR hasOutput a LE since the output of the process is a particular XMI file.

LE denotes a meaning. The meaning of the single token in the text is matched against the event type system in the SIMPLE database (i.e. the SIMPLE Ontology).

LE hasLinguisticAnnotation LA is carried out by the UIMA Type system, which is designed to accommodate both linguistic information, such as Part-of.Speech, Lemma and TimeML class.

The UIMA TS is a complex feature structure that is filled up during the process. When a potential event is encountered in the text it is tagged and the UIMA TS is filled.

Finally, LPR provides LS. LS are available from external users. Since this process provides only the TimeML tagging, the LS is only one.

Acknowledgements

We would like to thank the following people for their helpful contribution to the research activities towards to the realization of the Language Grid mission of an infrastructure of languages service ontology and composite language services: Francesca Bertagna, Riccardo Del Gratta, Valeria Quochi and Antonio Toral Ruiz.

5 References

Bertagna, F., Monachini, M., Soria, C., Calzolari, N., Huang, C., Hsieh, S., Marchetti, A., Tesconi, M., Fostering Intercultural Collaboration: a Web Service Architecture for Cross-Fertilization of Distributed Wordnets. In Ishida, T., Fussell, S. R., Vossen, P. (eds.) *Proceedings of the First International Workshop on Intercultural Collaboration (IWIC 2007)*, Kyoto, pp. 185-198. Also in: LNCS, vol. 4568, pp. 146-158. Springer, 2007.

Broeder, D. and Wittenburg P.. The IMDI metadata framework, its current application and future direction. IJMSO, 1(2):119–132. 2006.

Calzolari N., Towards a new generation of Language Resources in the Semantic Web vision. In Ahmad, K., Brewster, C., Stevenson, M. (eds.), *Words and Intelligence II: Essays in honour of Yorick Wilks*, pp. 63-105. Springer, 2007.

Calzolari N. Approaches towards a "Lexical Web": the role of Interoperability, In Proc2008 *ICGL2008*, Hong Kong, pp.34-42, 2008.

Calzolari, N., Bertagna, F., Lenci, A., Monachini, M. (eds.), *Standards and Best Practice for Multilingual Computational Lexicons. MILE (the Multilingual ISLE Lexical Entry).* ISLE, Pisa, 194 pp., 2003.

Caselli, T., Prodanof, I., Ruimy, N., Calzolari, N., Mapping SIMPLE and TimeML: improving event identification and classification using a semantic lexicon. In *GL2007: Fourth International Workshop on Generative Approaches to the Lexicon*. Paris, 2007.

Del Gratta R., Caselli T., Ruimy N., Calzolari N. TIME-ML: An ontological mapping onto UIMA Type System. In *ICGL2008*, Hong Kong, pp.105-112, 2008a.

Del Gratta R., Bartolini R., Caselli T., Monachini M., Soria C., Calzolari N. UFRA: A UIMA-based Approach to Federated Language Resource Architecture, In Proc. LREC2008 Marrakech, 2008b (to appear).

Fellbaum, C., Vossen. P., Connecting the Universal to the Specific: Towards the Global Grid. In Ishida, T., Fussell, S. R., Vossen, P. (eds.) *Proceedings of IWIC 2007*. Also in: LNCS, 2007.

Ferrucci, D., Lally, A., UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*. 10(3-4) 2004.

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M., Soria, C., Lexical Markup Framework (LMF). In *Proceedings of LREC2006*, Genova, pp. 233-236. ELRA, Paris, 2006.

Francopoulo , G., Declerck T., Sornlertlamvanich, V., De la Clergerie E., Monachini M., Data Category Registry: Morpho-syntactic and Syntactic Profiles, In Proc. of LREC2008. 2008 (to appear).

Hayashi, Y., Declerck, T., Buitelaar, P., and Monachini, M.. Ontologies for a Global Language Infrastructure. In: Proc. of *ICGL2008*, Hong Kong, pp.105-112, 2008a.

Hayashi, Y., Narawa, C., Monachini, M., Soria, C., and Calzolari, N.. Ontologizing Lexicon Access Functions based on a LMF-based Lexicon Taxonomy, In: Proc. of LREC2008. 2008b (to appear).

Ide, N., Calzolari, N., Introduction to the Special Inaugural Issue. *Language Resources and Evaluation*. Springer, 39(1), pp. 1-7, 2005.

Ishida, T., Language Grid: An Infrastructure for Intercultural Collaboration. In *IEEE/IPSJ Symposium on Applications and the Internet*, pp. 96-100, 2006.

ISO 24613 Lexical Markup Framework.

Marchetti, A., Tesconi, M., Ronzano, F., Rosella, M., Bertagna, F., Monachini, M., Soria, C., Calzolari, N., Huang, C.R., Hsieh, S.K., Towards an Architecture for the Global-WordNet Initiative. In *Proceedings of SWAP-06, 3rd Semantic Web Workshop*. 2006.

Monachini Monica. Test-suites of ISO conformant lexical entries. D2.3 LIRICS Deliverable, http://lirics.loria.fr/doc_pub/D2-3-noAppC.pdf, Pisa, 2007.

Monachini, M., Calzolari, N., Choukri, K., Friedrich, J., Maltese, G., Mammini, M., Odijk, J., Ulivieri, M., Unified Lexicon and Unified Morphosyntactic Specifications for Written and Spoken Italian. In *Proceedings of LREC2006*, Genova, pp. 1852-1857. ELRA, Paris, 2006.

Monachini, M., Quochi, V., Ruimy, N., Calzolari, N., Lexical Relations and Domain Knowledge: The BioLexicon Meets the Qualia Structure. In *GL2007: Fourth International Workshop on Generative Approaches to the Lexicon*. Paris, 2007.

Monachini , M., Quochi, V., Del Gratta R.., Calzolari, N., Using LMF to Shape a Lexicon for the Biomedical domain, In Proc. LangTech2008, Rome, 2008.

Pustejovsky J., José Castaño, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer and Graham Katz. 2003. TimeML: Robust Specification of Event and Temporal Expressions in Text. IWCS-5, Fifth International Workshop on Computational Semantics.

Quochi, V., Del Gratta, R., Sassolini, E., Monachini, M., Calzolari, N., Toward a Standard Lexical Resource in the Bio Domain. In Vetulani, Z. (ed.), *Proceedings of 3rd Language and Technology Conference*, Poznań, pp. 295-299, 2007.

Soria, C., Tesconi, M., Marchetti, A., Bertagna, F., Monachini, M., Huang, C., Calzolari, N., Towards agent-based cross-lingual interoperability of distributed lexical resources. In Proceedings *of COLING-ACL Workshop Multilingual Lexical Resources and Interoperability*, Sydney, 2006.

Tokunaga, T., Sornlertlamvanich, V., Charoenporn, T., Calzolari, N., Monachini, M., Soria, C., Huang, C., Prevot, L., Xia, Y., Yu, H., Kiyoaki, S., Infrastructure for standardization of Asian language resources. In *Proceedings of COLING/ACL 2006 Main Conference Poster Sessions*, pp. 827-834. Sydney, 2006.

Tokunaga, T., Kaplan D., Sornlertlamvanich, V., Charoenporn, T., Calzolari, N., Monachini, M., Soria, C., Huang, S.K. Hsiech.,., Kiyoaki, S. YingJu X., Adapting International Standard for Asian language technologies. In *Proc LREC2008*, 2008 (to appear).

Toral, A., Monachini, M., Formalising and bottom-up enriching the ontology of a Generative Lexicon. In *Proceedings of RANLP07 - Recent Advances in Natural Language Processing*. Borovets, Bulgaria, 2007.

Van Assem, M., Gangemi, A., Schreiber, G., Conversion of WordNet to a standard RDF/OWL representation. In *Proceedings of LREC2006*, Genova. ELRA, Paris, 2006.